Next Token Generation & Word Embeddings

Nipun Batra

IIT Gandhinagar

July 31, 2025

Outline

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Acknowledgment & Inspiration

Inspired by Excellence

This presentation is inspired by the excellent lecture from **Andrej Karpathy**

Find the original lecture:

- Search for: "Neural Networks: Zero to Hero"
- Andrej Karpathy's educational content

Relevance to Modern AI

Connection to ChatGPT: The same fundamental principles we'll learn here power modern language models like GPT-4!

The Core Question

app

What is the next character?

Next Character Prediction

app?

Next Character Prediction

app?

Pose as Classification Task

We can frame this as a **multi-class classification problem**

Next Character Prediction

app?

Pose as Classification Task

We can frame this as a **multi-class classification problem**

Next Character Prediction

app?

Pose as Classification Task

We can frame this as a **multi-class classification problem**

Next Character Prediction

app?

Pose as Classification Task

We can frame this as a **multi-class classification problem**

Input:

app

Next Character Prediction

app?

Pose as Classification Task

We can frame this as a **multi-class classification problem**

Input:

app

Output: Probability Distribution

Character	Probability
a	0.01
b	0.01
c	0.03
...	...
l	**0.85**
...	...

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

a) Text is too long for neural networks

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

a) Text is too long for neural networks
b) Neural networks only work with numerical inputs

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Pop Quiz: Text Representation

Quick Quiz 1

Why can't we directly feed text into neural networks?

a) Text is too long for neural networks
b) Neural networks only work with numerical inputs
c) Text doesn't contain useful information

Answer: b) Neural networks perform mathematical operations requiring numerical inputs!

Dataset: Indian Names

Our Training Dataset

Dataset: Indian Names

Our Training Dataset

Dataset: Indian Names

Our Training Dataset

- **Abid**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮
- **Priya**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮

- **Priya**
- **Rahul**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮
- **Priya**
- **Rahul**
- **Sita**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮

- **Priya**
- **Rahul**
- **Sita**
- **Vikram**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮

- **Priya**
- **Rahul**
- **Sita**
- **Vikram**
- ⋮

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮
- **Priya**
- **Rahul**
- **Sita**
- **Vikram**
- ⋮
- **Zara**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮
- **Priya**
- **Rahul**
- **Sita**
- **Vikram**
- ⋮
- **Zara**

Dataset: Indian Names

Our Training Dataset

- **Abid**
- **Abhidha**
- **Adesh**
- **Ajay**
- **Akash**
- **Ananya**
- ⋮
- **Priya**
- **Rahul**
- **Sita**
- **Vikram**
- ⋮
- **Zara**

Goal

Learn to generate new, realistic Indian names character by character

Dataset Assumptions

Simplifying Assumptions

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)
2. **End marker:** Hyphen (-) indicates end of name

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)
2. **End marker:** Hyphen (-) indicates end of name

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)
2. **End marker:** Hyphen (-) indicates end of name
3. **Length:** Names are between 4-10 characters

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)
2. **End marker:** Hyphen (-) indicates end of name
3. **Length:** Names are between 4-10 characters

Dataset Assumptions

Simplifying Assumptions

1. **Alphabet:** Only 26 lowercase characters (a-z)
2. **End marker:** Hyphen (-) indicates end of name
3. **Length:** Names are between 4-10 characters

Example Encoding

"Abid" becomes **"abid-"**

Generate Training Dataset

From one name, create multiple training examples

Generate Training Dataset

From one name, create multiple training examples

Example: "Abid" → "abid-"

Generate Training Dataset

From one name, create multiple training examples

Example: "Abid" → "abid-"

Using Context/History of 3 Characters

X (Input)	Y (Target)
...	a
..a	b
.ab	i
abi	d
bid	-

Generate Training Dataset

From one name, create multiple training examples

Example: "Abid" → "abid-"

Using Context/History of 3 Characters

X (Input)	Y (Target)
...	a
..a	b
.ab	i
abi	d
bid	-

Generate Training Dataset

From one name, create multiple training examples

Example: "Abid" → "abid-"

Using Context/History of 3 Characters

X (Input)	Y (Target)
...	a
..a	b
.ab	i
abi	d
bid	-

Result: 5 training examples from one name "abid"!